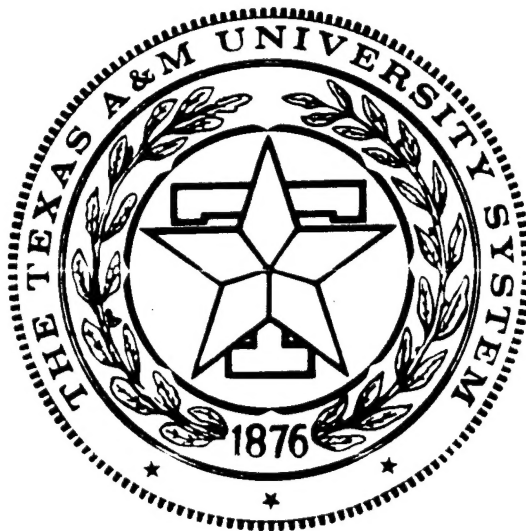TEXAS A&M UNIVERSITY

Hardware Implementation of a
Desktop Supercomputer for
High Performance Image Processing

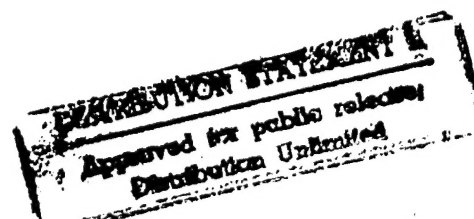ONR Grant Number N00014–94–1–0516

Technical Report
Feb/01/95 – May/01/95

DEPARTMENT OF ELECTRICAL ENGINEERING
College Station, Texas

Hardware Implementation of a Desktop Supercomputer for High Performance Image Processing

Technical Report
Feb/01/95 – May/01/95

CNN VLSI IMPLEMENTATION

Dr. Jose Pineda de Gyvez

Texas A&M University
Microelectronics Group
Department of Electrical Engineering
College Station, TX, 77843

Phone: (409) 8457477
FAX: (409) 8457161
Email:gyvez@pineda.tamu.edu

TABLE OF CONTENTS

I. INTRODUCTION

CNN is a spatio–temporal first order non–linear filter[1] that can be expressed as

$$C\frac{dx_{i,j}(t)}{dt} = -\frac{1}{R}x_{i,j}(t) + \sum_{(k,l)\in N_r(i,j)} A(i,j;k,l)y_{k,l}(t) + \sum_{(k,l)\in N_r(i,j)} B(i,j;k,l)u_{k,l}(t) + I$$

$$y_{i,j}(t) = \frac{1}{2}(|x_{i,j}(t) + 1| + |x_{i,j}(t) - 1|)$$

where **A** and **B** are parameters matrices of size N_r, and I is a constant. These coefficients determine the function of the filter. u_{ij} is a two dimensional constant signal while x_{ij} is a time varying state variable of a cell. The input signal can be applied as either *X(0)* or *U,* that is a matrix of { x_{ij} } and { u_{ij} }, respectively. The output signal is taken from either x_{ij} or y_{ij}. The neighborhood size is typically 3x3 in most applications even though several reported applications require larger size. In hardware implementations, larger neighborhood sizes are very costly due to the complex interconnections. Figure 1 shows the connections for one cell that has a 3x3 neighborhood.

Even though each cell is locally connected, the hardware implementation of an array that has a practical size is extremely difficult. Even the most advanced 'state–of–art' implementation has only 32x32 cells which is far below any practical size[2]. Multiplexing is unavoidable for large images in current technologies until any significant break–through of technology is realized. A modular approach is one possible solution but the parasitic capacitance of the interconnection may cause some problems. Moreover an advanced packaging technology is required for modular approach.

Since the objective of this project is to construct a multiplexing CNN processor, the circuit and architecture should be suitable for multiplexing. The optimal size of the processor is about 40x40 cells considering a typical image size to be processed, expected feature size, available die size, number of available pins, and total processing time including data interface. These observa-
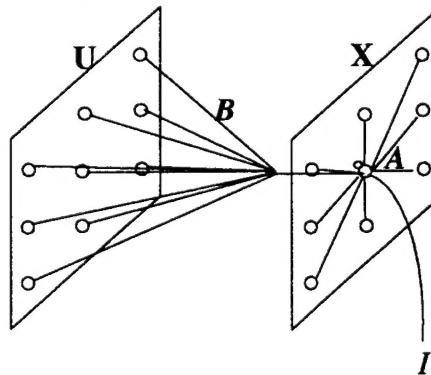

Figure 1. Cell connections

tions give a outline of design constraints stated as '40x40 cells on 1cm^2 die'. Even though this processor is not supposed to be used in portable equipment, the power consumption must be minimized to implement a large array.

The proposed design is fully programmable providing the capability to adjust time constant and template range. The circuit is designed using mixed–mode design technology for optimization. The I/O interface is performed by a Switched–Capacitor circuit, the multiplier is a transconductance multiplier and the integration is performed by OTA–C circuit.

Detailed circuit design issues are discussed in the following section. Layout issues are discussed in section 3.

II. CIRCUIT DESIGN

Each cell in the CNN array has 18 multipliers, one lossy integrator, one hard limiter and one sample/hold to retain the u variable. Since CNN consists of a large array of cells, the circuits need to be of low power consumption and need to occupy a small area. Concurrently, they need to have reasonable computational accuracy to minimize the accumulation of errors over a large array. Even though this error may not affect the performance for certain applications, it needs to be minimized to use the implemented hardware as a universal CNN array. Fig. 2 shows the block diagram of one cell. The multipliers are divided into two groups; one for **A** template and the other for **B** template. All multipliers in a same group share one input which is output of hard limiter and Sample–Hold, respectively. The multiplier is a transconductance multiplier, so the summation is performed by one node in front of the integrator. A detailed circuit design method for each
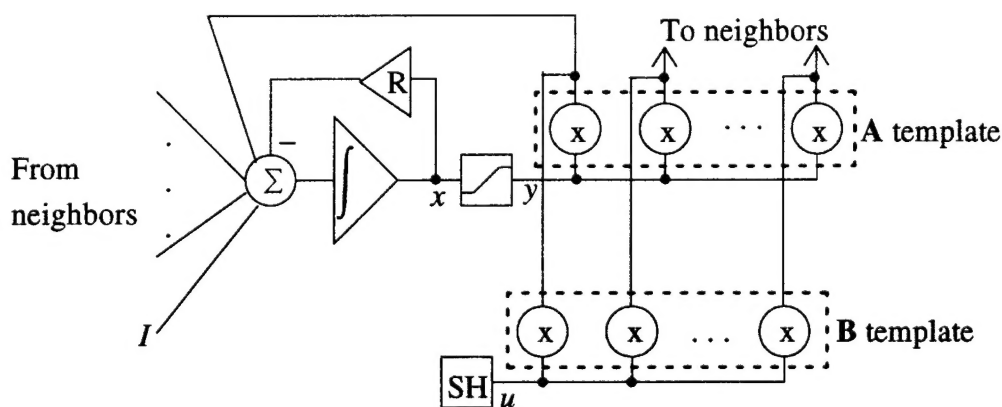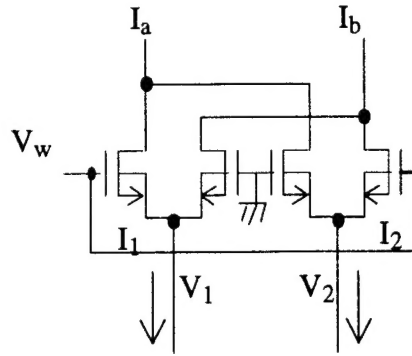

Figure 2. Cell Block Diagram

Figure 3. Basic multiplier building block

building block that is discussed in the following subsections. The complete circuit diagram of a cell, which was sent for test fabrication through MOSIS, is shown in Appendix B.

II-1. MULTIPLIER

Two cross coupled differential pairs in Fig 3 generate a differential current when all the transistors are operating in the saturation region as follows

$$I_{out} = KV_w(I_1 - I_2) = K'V_w(V_1 - V_2)$$

The previous equation suggests two possible signal injection methods. One is voltage–voltage and the other is current–voltage. From the analysis, it turns out that the signals applied to the gates need not necessary to be fully differential as long as the λ–effect of the MOS–FET is negligible. However (I_1, I_2) or (V_1, V_2) have to be fully differential to keep the symmetry. Several different output modes are possible with the multiplier described above. The output can be either fully–differential or single–ended and be either voltage or current.

A multiplier with voltage inputs and fully–differential voltage output is shown in Fig. 4[3]. The differential amplifier at the input stage provides the common mode bias voltage for the multiplier. This differential amplifier allows a single ended input without causing significant symmetry degradation. The multiplier output current is injected to the folded cascode circuit. A common mode feedback circuit is provided to fix the common mode voltage. A differential current output can be also obtained if additional N–FET cascode transistors (encircled by dotted line) are provided to increase the output impedance.
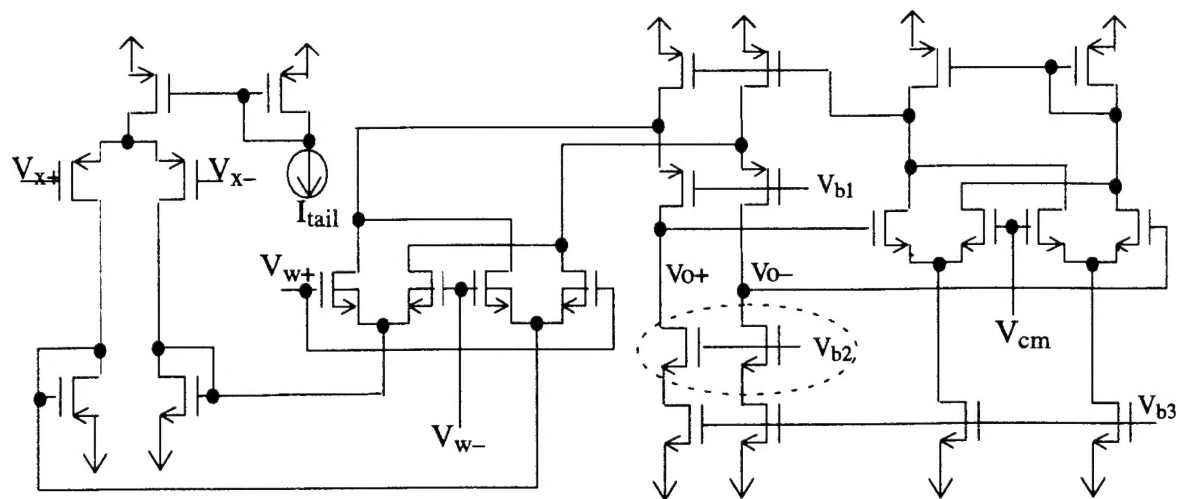
3

Figure 4. Voltage multiplier

Fig. 5 shows the measured characteristic of the fabricated multiplier in Fig 5. The power consumption is 200µW with a power supply voltage of ±3V. The area occupied is 0.06mm². The summation can be performed by cascading the multiplier as in Fig. 6. This summation scheme makes the 16 inter–cell connection inevitable. This requires large area and the routing becomes very complicated.

If the output of the multiplier is a single ended current signal then the summation can be performed by connecting the outputs of the multipliers together. As a result, the required number
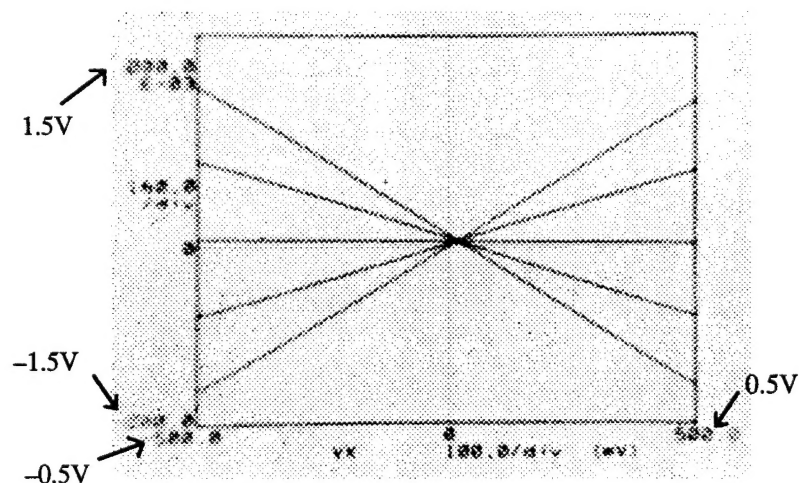

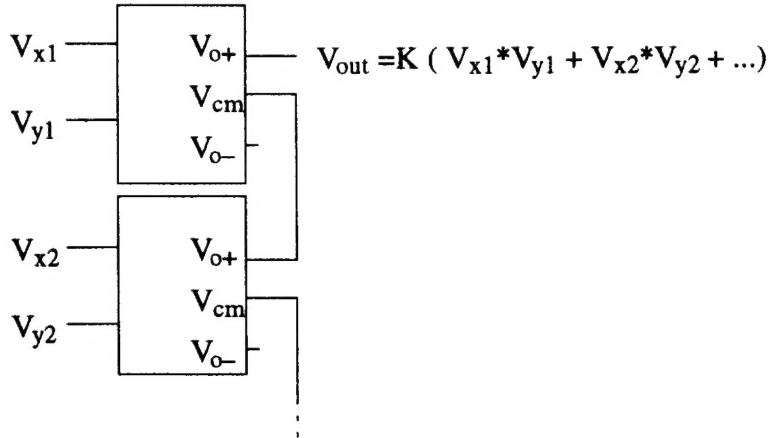Figure 5. Measured multiplier characteristic

4

Figure 6. Sum–of –Product computation
using voltage multipliers

of inter–cell connections is reduced to 8 and the routing becomes very simple. A detailed routing scheme will be discussed in following LAYOUT section.

Since the multiplier in Fig 3 is a current mode multiplier, the single–ended current mode output can be obtained by providing a current mirror on top of the multiplier as in Fig 7.

The fully differential input current is provided by a hard limiter which is a differential amplifier(a detailed discussion is provided in following subsection). This multiplier can be considered a modification of the Gilbert Multiplier[4]. The main difference is the injection of the differ-
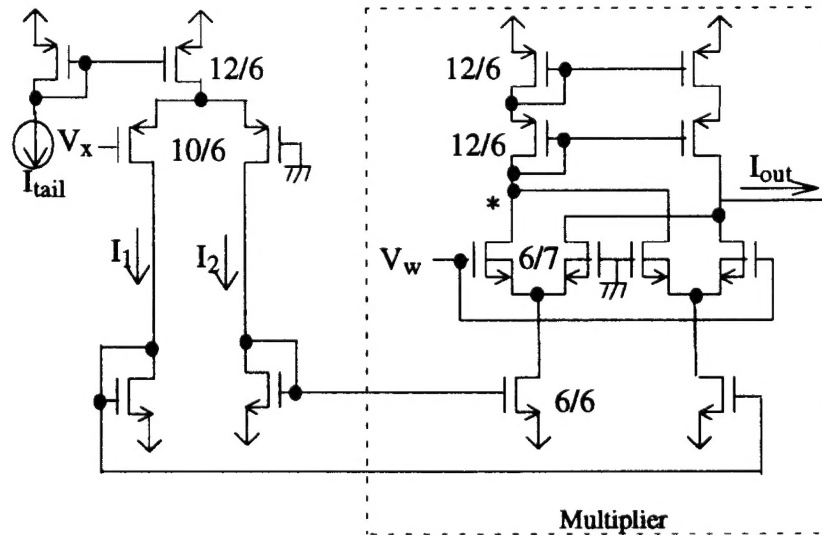

Figure 7. Transconductance multiplier

5

ential input current. The Gilbert multiplier implements a differential pair under the multiplier of Fig 3. Hence, two transistors are attached under the multiplier. This multiple stacked transistor scheme requires high power supply voltages. As several transistors are stacked, the voltage at the input node of the multiplier in Fig 3 varies widely causing high λ-effects. As a result, the actual linear dynamic range in the Gilbert cell is very small even with high power supply voltages. The proposed design uses a current mirror to avoid stacking the transistors so that the linear dynamic range is improved. In the Gilbert multiplier V_w and V_x cannot be single ended by fixing one input to ground because the transistor would be not properly biased. Two separated bias voltages have to be applied to the fixed input which means additional connections.

The V_w signal can be a single ended voltage. Even though the single ended unbalanced input causes non–symmetricity due to the λ-effect of the MOS FET, its effect is negligible. Since the weight signal has to be connected to all the cells in the array, a single ended voltage signal reduces the number of connections. The voltage difference between the marked (*) node and the output node causes serious degradations of symmetricity due to the λ-effect of the MOSFET. Since the output stage requires a high output impedance, the voltage at the output node varies widely as the output current varies. This effect can be minimized by connecting the multiplier's output node to the negative terminal of an OTA with negative feedback. This will be discussed in the following subsection. While the output node is fixed to a virtual ground, the marked node should have the same operating voltage to minimize non–symmetric effects. This is achieved by properly selecting the size of the transistors of the current mirror.

Fig. 8 shows simulation results of the designed multiplier with the hard limiter. The tail current of the differential pair is set to 9µA and the power supply voltage to ±3 Volts. The linear dynamic range of V_w and V_x is constrained by the tail current, and the W/L ratios of the transistors in the differential pair and the multiplier. The linear dynamic range of V_w and V_x are designed to be ± 0.5V with maximum differential output current of ±6µA. This is a the trade off between power consumption, area and dynamic range. Actually, the required dynamic range is smaller due to the dynamic range limitation of the cell state. In the actual design, which was sent for fabrication, the bias current is externally controlled so that the linear range can be programmable. The power consumption of this multiplier is 54µW. The output impedance is 5MΩ for $V_w = 0.2$ Volt and 4.5µA of differential current. The occupied area is 0.008mm^2 in 2µ Orbit technology with the N–well process.
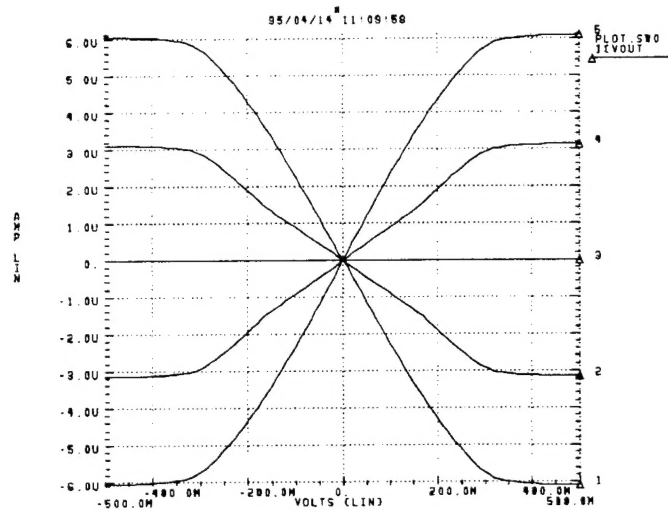
6

Figure 8. Simulation results of transconductance multiplier

II–2. HARD LIMITER

Since the integrator has a single–ended voltage output while the multiplier requires fully differential current inputs, a differential pair is used as a hard limiter to generate a differential current. These currents are copied to all multipliers belonging to the **A** template by current mirrors as illustrated in Fig 9. The linear range is controlled by the tail current with the given transistor sizes of differential pair. The linear range of the hard limiter is designed to be ±0.5V with 9μA of tail current. Fig. 10 shows the programmability of the differential pair as a hard limiter.
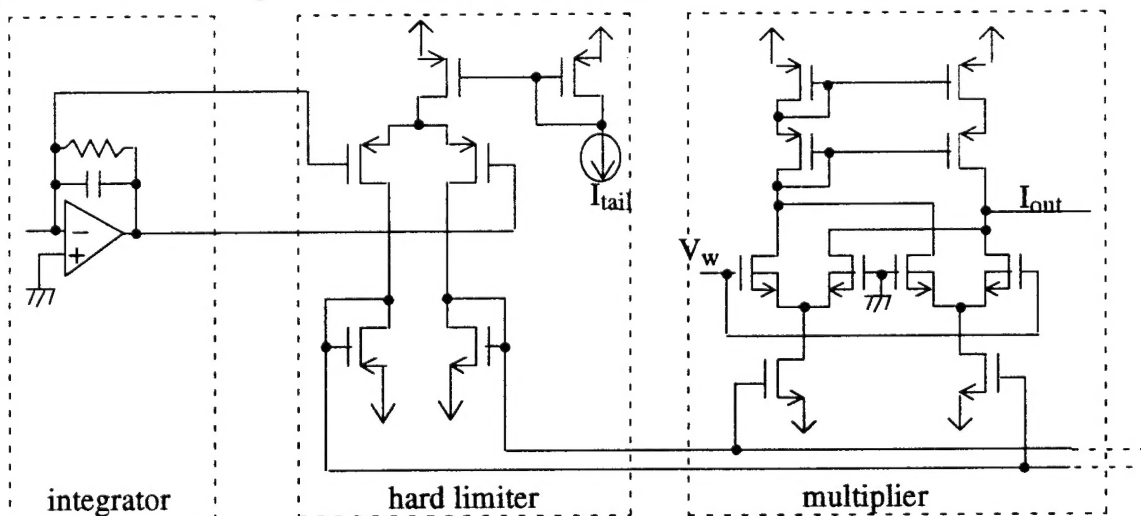

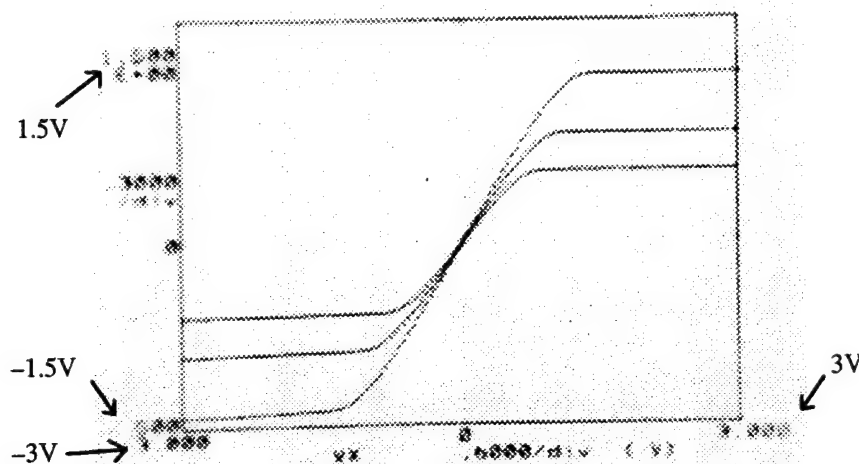Figure 9. Hard limiter and connections to other building blocks

Figure 10. Measured programmability of the hard limiter

Since the output of the OTA integrator is single ended, The other input of the differential pair is connected to the inverting input of the OTA to eliminate the effect of the offset voltage of the OTA. The linear range of the hard limiter needs to be programmable.

II–3. SAMPLE/HOLD

A differential pair is used to convert u_{ij} signals that are stored as a voltage in a holding capacitor into a fully differential current signal. A 1pF capacitor is provided at the input of the differential pair to hold the u_{ij} signal. The clock feedthrough is quite serious in this configuration. The clock feedthrough can be cancelled with a fully differential configuration. If an additional holding capacitor(dotted in Fig. 11) is provided at the other gate of the differential pair, and the input signal is fully differential then most of the feedthrough effect can be cancelled out. If the input
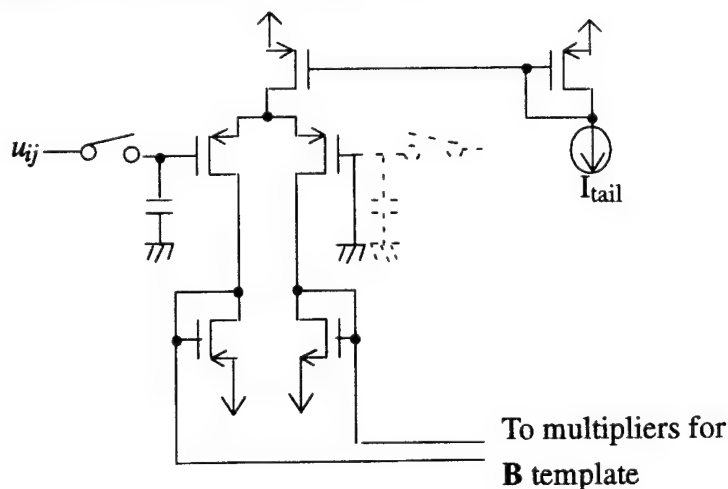

Figure 11. Sample/Hold

8

image is binary then the feedthrough effect does not degrade the image. Since the input scheme for x_{ij} also has signal degradation, the actual feedthrough effect is compensated a little bit. For simplicity, only one holding capacitor was implemented with a single–ended configuration. The differential amplifier is the same as the hard limiter.

II–4. RESISTOR

The R in equation (1) has to satisfy $A(i,j;i,j) > \frac{1}{R}$ to ensure convergence. It should satisfy $V_{x\,max} > I_{in\,max} \times R$ where I_{in} is the total current injected into the integrator to prevent the saturation of V_x. If we map an analytical 1 of a template to $V_w = 0.1V$ in the actual circuit, then the transconductance of the multiplier at $V_w = 0.1V$ represents $A(i,j;i,j)$ in the actual circuit. Since the transconductance of the designed multiplier at $V_w = 0.1V$ is 6/5 μA/V, R has to be greater than 5/6MΩ. This resistance occupies unreasonably large silicon area when implemented through a typical defusion resistor.

The resistor can be implemented with a MOSFET (Fig. 12) operating in the ohmic region but it's linearity is very poor besides the dynamic range is small. The linearity and dynamic range can be improved with a fully differential configuration but it is not acceptable because we need to use a single ended circuit to minimize the number of inter–cell connections. Another possibility is to use an extra power supply that can be applied only to the gate of the transistors. However, if V_{gs} is high then the conductance of the transistor is increased. As a result, the length of the transistor needs to be unusually long. Since the effect of extremely long transistor lengths is not practical this has to be considered carefully.

The resistor can be implemented with an OTA as in Fig. 13. This active resistor implementation has better linearity than the MOSFET resistor, but it has two disadvantages. The first is a small dynamic range and the second one is the offset. The linear range may be increased with a linearized differential pair. This will be studied further. The effect of the offset in the cell can be cancelled out by adjusting the bias current (I_{tail}). Due to the limited dynamic range of this OTA resistor, the power supply voltage needs to be increased to ±3V even though all other cir-
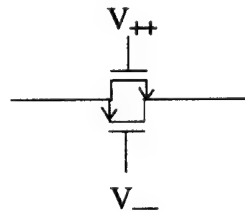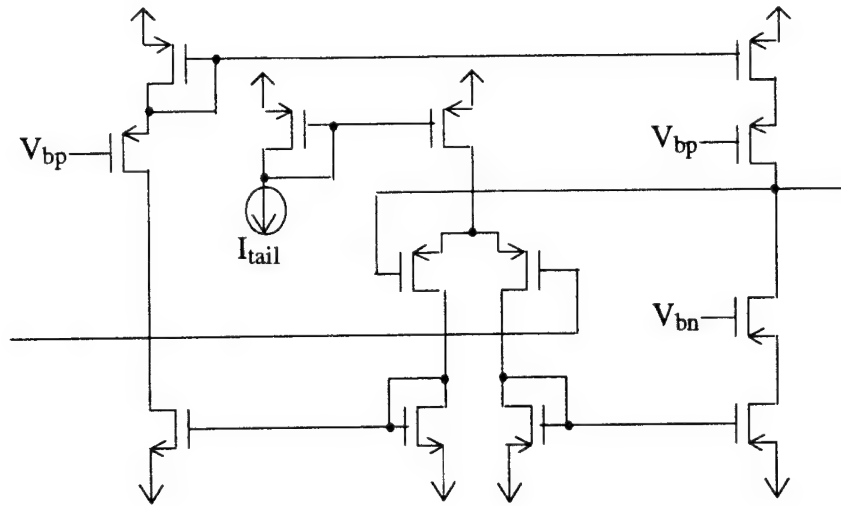

Figure 12. MOS FET resistor

Figure 13. OTA resistor

cuits can operate with a ± 2 V power supply. The designed OTA resistor has ± 2 V of dynamic range and consumes 600 μW.

This approach also increases the power consumption by a factor of 50% due to the increased power supply voltage. The OTA consumes 11 times more power than the multiplier. This means that we can implement the multiplier with a higher transconductance resulting in a lower required resistance. Since the amount of area increased for a higher transconductance is dependent on the layout while the area of the diffusion resistor is dependent on the fabrication technology, the selection of the resistor implementation method has to be examined more carefully to optimize the design.

A resistor can be implemented with the transconductance multiplier shown in Fig 7. The linear range is the same as that of the OTA resistor. The main advantage of this multiplier–resistor is the programmability that allows some optimization in mapping analytical template values into circuit signals. The transistor sizes are different from the multiplier in order to increase the dynamic range. The designed multiplier–resistor consumes 324μW with ± 3V power supply, occupies 0.01mm^2 in 2μ N–well technology, has a dynamic range of ± 1.3V and is programmable from 250KΩ to infinity. In a P–well process, the required area is a little bit smaller and the dynamic range can be increased to ± 1.8V with the same power consumption. The simulated result is shown in Fig. 14.
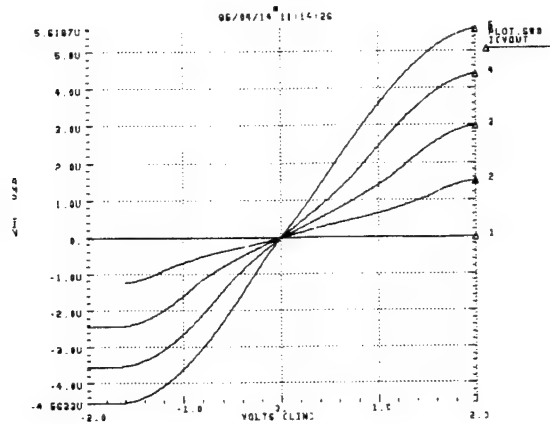
10

Figure 14. Simulation result of a programmable resistor

II–5. BIAS

The bias (I at Fig.1 and Eq. (1)) circuit can be implemented with a simple current mirror as shown in Fig. 15. If an OTA resistor is used then the bias voltage for the cascode stage can be shared with the OTA. Only one input stage of the current mirror is provided on chip and two gate nodes are connected to all cells in the chip.

The above approach requires two connections for the mirror and two more additional bias connections for the cascode. A single–ended voltage–controlled current source is implemented with a telescope OTA as in Fig 16.
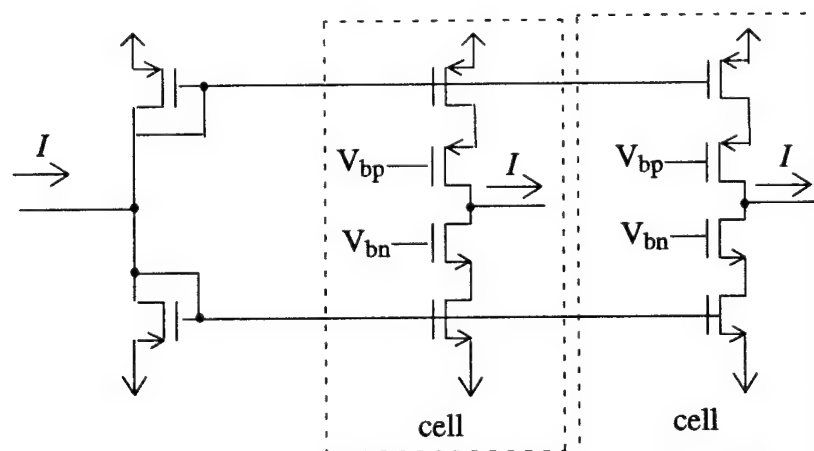

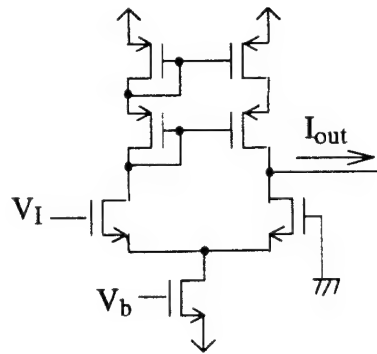Figure 15. Bias current copier distribution

11

Figure 16. Bias circuit

II-6. INTEGRATOR and OTA

We considered two possible implementations of current lossy integrators. The first method consists of a parallel connection of a resistor and a capacitor as shown in Fig. 17(a). The second method consists of an integrator with an OTA as in Fig 17(c). Since the outputs of 18 multipliers and the bias circuit are connected together to achieve the current summation, the equivalent output resistance of the multipliers and the bias circuit are reduced by a factor of 20 (200KΩ). The sum of the current is not fully transferred into the integrator due to this loading effect. If the output node of a multiplier is left floating then the symmetry is degraded because the voltage of marked (*) node in Fig. 7, is practically fixed due to the diode connected transistor in the current mirror. As a result, the ratio of the current that is transferred from the multipliers into the integrator is dependent on the voltage at the summing node. One possible solution for this effect is to separate the summing node from the integrating node with a current mirror as in Fig 17 (b). Even though there is a small voltage variation at the input node, the operating point is highly dependent on the process variation. This circuit has no control of output offset voltage and current and therefore will degrade the accuracy seriously. This effect can be minimized using feedback as in Fig. 17(c). The feedback loop keeps the summing node to a fixed voltage.
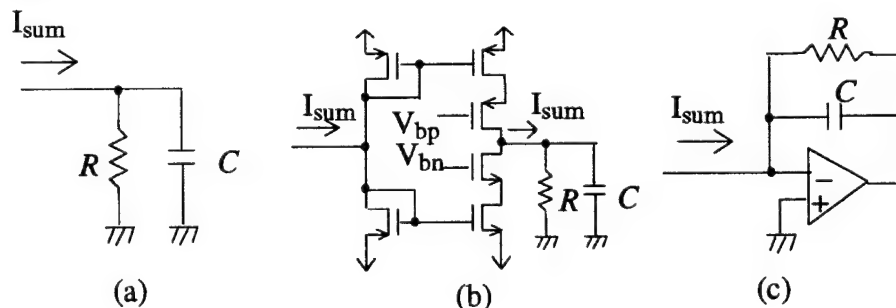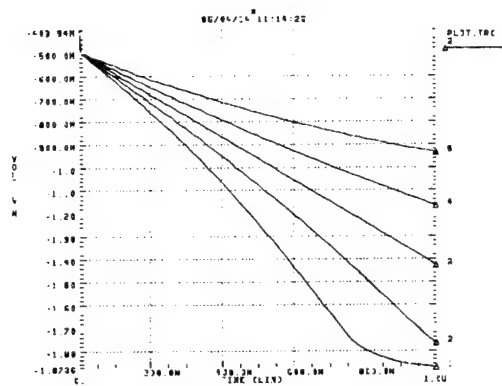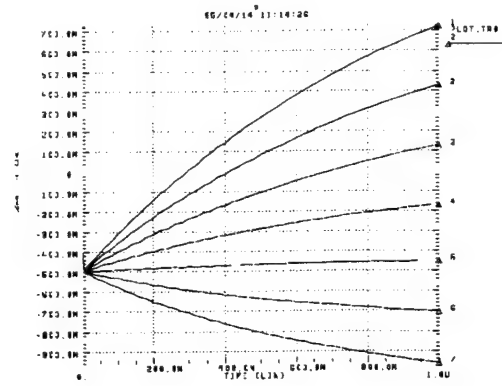

(a)   (b)   (c)

Figure 17. Lossy integrator

12

(a) Various time constant   (b) various input current

Figure 18. Simulation result of OTA integrator

If the resistor is implemented with an OTA–resistor or a Multiplier resistor then the output node of the OTA is connected to the gate of resistor. This connection does not drain any current from the OTA. As a consequence, the OTA drives the capacitor only. Since OTAs are easier to design with high gain and high GB than conventional OP Amps for fixed area and power consumption, the OTA is used instead of the OP Amp. Fig. 18(a) shows simulation results with various time constants using an initial condition of –0.5V, and a 2μA constant input current. Fig. 18(b) shows simulation results for various input currents with a fixed time constant.

The tail current of the OTA has to be greater than the maximum input current to sink or source the required amount of current. The designed OTA consumes 720μW and occupies 0.017mm². It has 40 db DC gain, 70Mhz GB and 70° phase margin.

II–7. I/O AND CONTROLLING CIRCUITRY: CORE

The core is a collection of switches to read/write the data and to control the dynamics as illustrated in Fig. 19. The switches are implemented using transmission gates. A transmission gate has three main non idealities. The first is the feedthrough, the second is the finite OFF resistance and the third is the ON –resistance. The effect of ON–resistance is not significant in the designed configuration.

The SELECT switch selects a cell to transfer the data. Since the SELECT switch forms a bilinear configuration, the effect of the feedthrough is cancelled out. After one cell is selected the state of the cell is read out during the clock phase ϕ_1. At the same time the initial state for
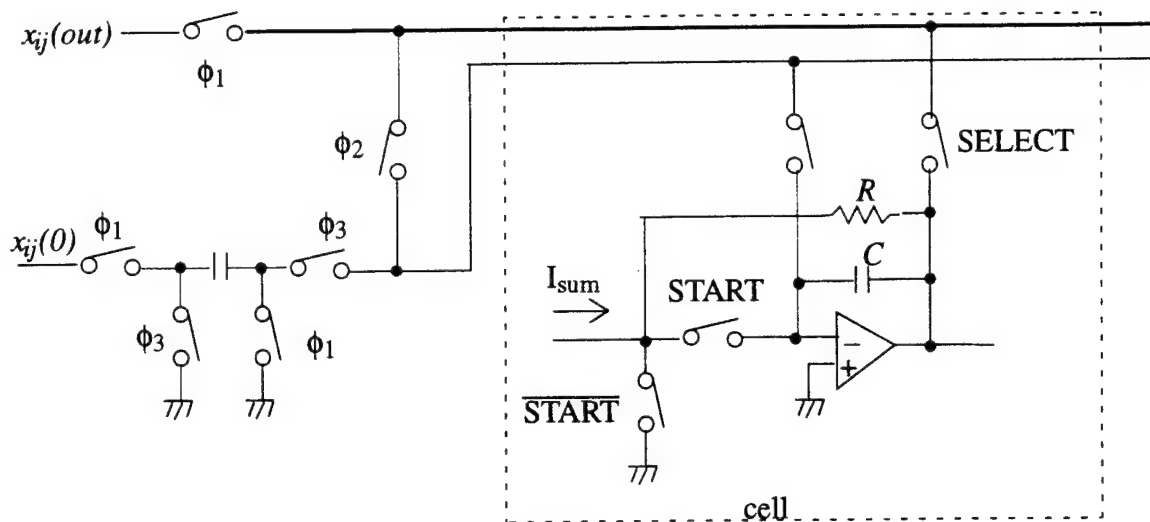
Figure 19. Core

the next processing image block is transferred into the buffer capacitor. The capacitor in the integrator is discharged during the ϕ_2. The charge in the buffer capacitor is transferred during ϕ_3.

During the data transfer stage the START switch is opened at ϕ_2 to stop the dynamics. $\overline{START}$ switch is provided to drain the incoming current to minimize the leakage current while the START switch is off. The START switch has to be kept ON after the dynamics have started until the data is read out. If this switch is turned off after all the cells have converged, the feedthrough degrades the data in the capacitor. This switch has to be turned off before the new initial state is transferred. Since the amount of feedthrough is dependent on the signal voltage, it causes serious nonlinear offset. This makes the switch control complicated because the START switch in each cell has to be controlled separately which requires a flip–flop in each cell. This non–ideality may not serious when only the hard limited output is required. Since feedthrough is more serious for high signal voltages, its effect after hard limiting becomes relatively small.

Another clocking scheme with the given switch configuration is possible in order to solve this problem. Once the array has converged, the state of the cells are read out one by one while keeping the START switches for all cells closed. Once all the data is read, the START switch is turned off and then the cells are selected one by one again to set the new initial state. The SELECT, ϕ_1 and ϕ_2 signals are turned on at the same time, and then the initial condition is transferred at ϕ_3.

Even tough the cell is accessed twice in the latter scheme the total data transfer time for both schemes is identical. The possible disadvantage of the latter scheme is that it requires a buffer

14

memory for the whole image if the input data is a stream of frames. The former scheme is suitable for pipelinning processes because the input and output sequences are synchronized.

Since the data transfer is performed by a Switched–Capacitor technique, the speed of interface is limited by the finite GB and slew rate of the OP Amp in the integrator.

III. LAYOUT

One of the most critical issues in the CNN array design is silicon area. The area has to be minimized. The following layout scheme is optimal for our architecture. Even though the detailed layout is dependent on the circuit design, the proposed layout scheme may also be applied to various circuit implementations.

The array of multipliers occupies almost 2/3 of the cell's area. One cell has 18 multipliers, thus 18 weight signals have to pass over all cells in the array. The circuit uses metal 1 with the power lines running horizontally as in digital VLSI. Metal 2 can be used to pass the weight signal vertically. Two multipliers whose outputs have to be connected to the same cell form a pair so that they minimize the summing line as in Figure 20. The multiplier array can be build up by putting 9 multiplier pairs together. Even though the connection from the hard limiter to the multipliers is fully differential, it does not cause complex connections because all the multipliers are close to each other and to the hard–limiter. Figure 21 shows a floor plan and corresponding routing plan for a cell. Fig.22 shows the layout of a multiplier
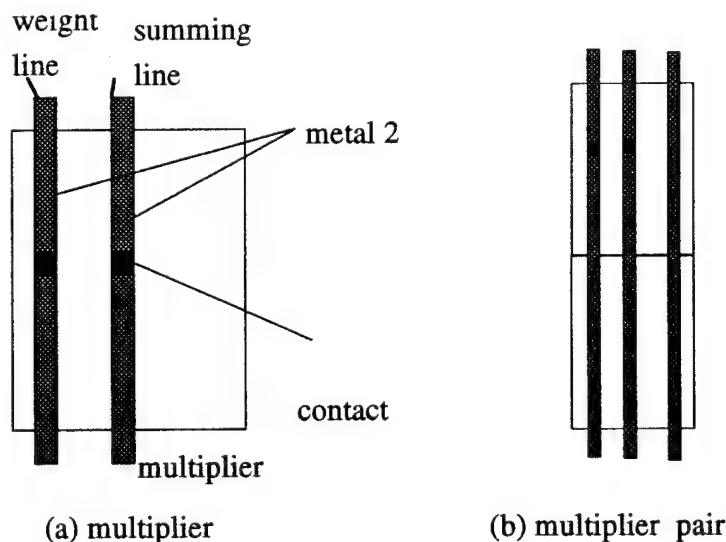

(a) multiplier   (b) multiplier pair
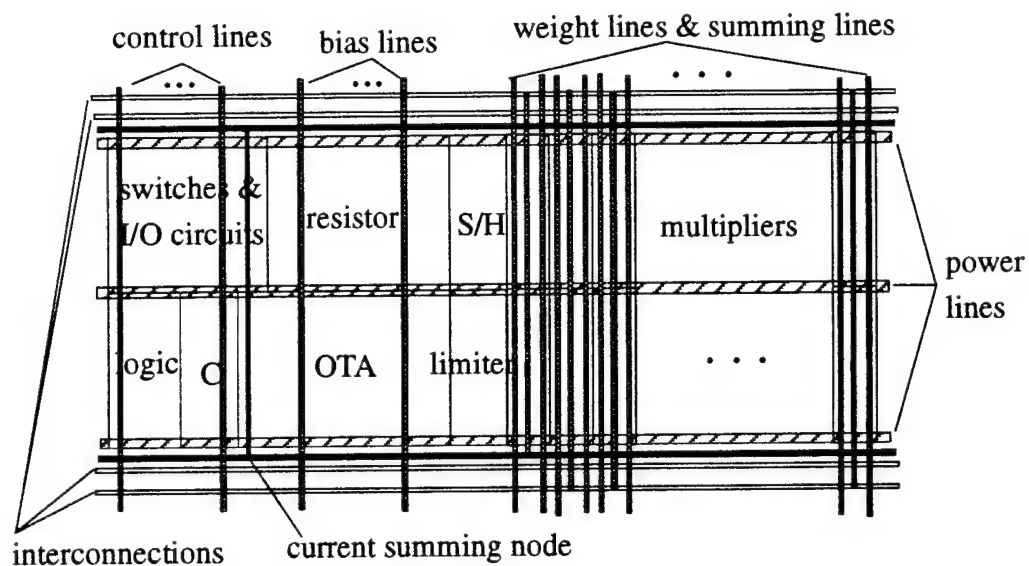
Figure 20. weight and summing line
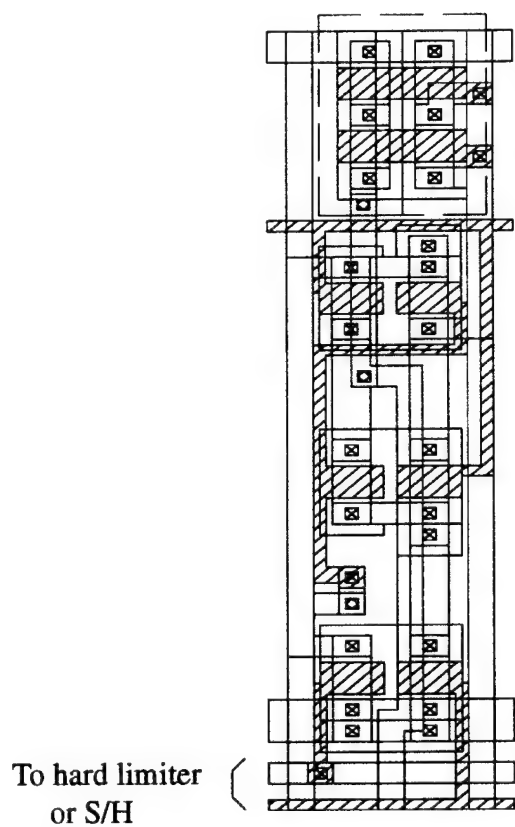
Figure 21. Floor plan

Figure 22. Layout of multiplier

IV. CONCLUSIONS AND FURTHER PLAN

The summary of characteristics of the designed building blocks that were sent for test fabrication are shown in Appendix A. The designed cell is fully programmable, consumes small power (2.5mW/cell) and may have reasonable accuracy. Since this cell consumes small power and occupies small area ($0.25mm^2$), it is suitable for very large array implementations.

Two chips that have basic building blocks were sent for fabrication on March 8 which corresponded the P–well process. One chip that has one cell excluding interconnection lines was sent for fabrication on April 5 which was the N–well process. This cell includes some I/O circuitry even though the final I/O scheme is not completely decided.

Further research in power consumption reduction and accuracy improvement will be performed. After testing the fabricated building blocks, a 4x4 array on a tiny chip will be sent for fabrication. Intensive research will be performed on the scaling and on the non–ideality effects of circuits will be performed. Once these chips are tested, a larger array (possibly 15×15 or larger on 6.9 x6.9mm^2 die) will be sent for fabrication.

REFERENCES

[1] Leon O. Chua, *et al*, "Cellular Neural Networks: Theory", *IEEE Trans. on Circuits & Systems*, vol. 35, no. 10, Oct. 1988, pp1257–1272.

[2] Espejo S., *et al*, "A 1μm CMOS cellular neural network universal machine", *Proc. of IEEE International Workshop on Cellular Neural Networks & their Applications*, pp 91–96.

[3] Han G.,*et al*, "A general purpose discrete–time multiplexing neuron array architecture", *Proc. of ISCAS,* vol. 2, May 1995, pp1320–1323.

[4] B. Gilbert, "A high–performance monolithic multiplier using active feedback", *IEEE J. Solid–State Circuits,* vol. SC–9, Dec. 1974, pp364–373.

Appendix A: Summary of the designed building blocks

Building Block	Type	Mode	Power Consumtion	Area	Dynamic range	Programmability
Multiplier	Trans–conductance	Input : single–ended differential Output : single–ended	54μW	0.008mm^2	Programmable	scalable template range
Resistor	Multiplier		324μW	0.01mm^2	±1.2V	250KΩ ~ infinity
Integrator	OTA integrator	Single–ended	720μW	0.017mm^2	±1.2V	variable time constant
hard limiter	Differential Amp	Input : single–ended Output : differential	54μW	0.008mm^2	program–mable	dynamic range
I/O	Switched–Capacitor	Single–ended				I/O schedule

Appendix B: Circuit diagram of a cell
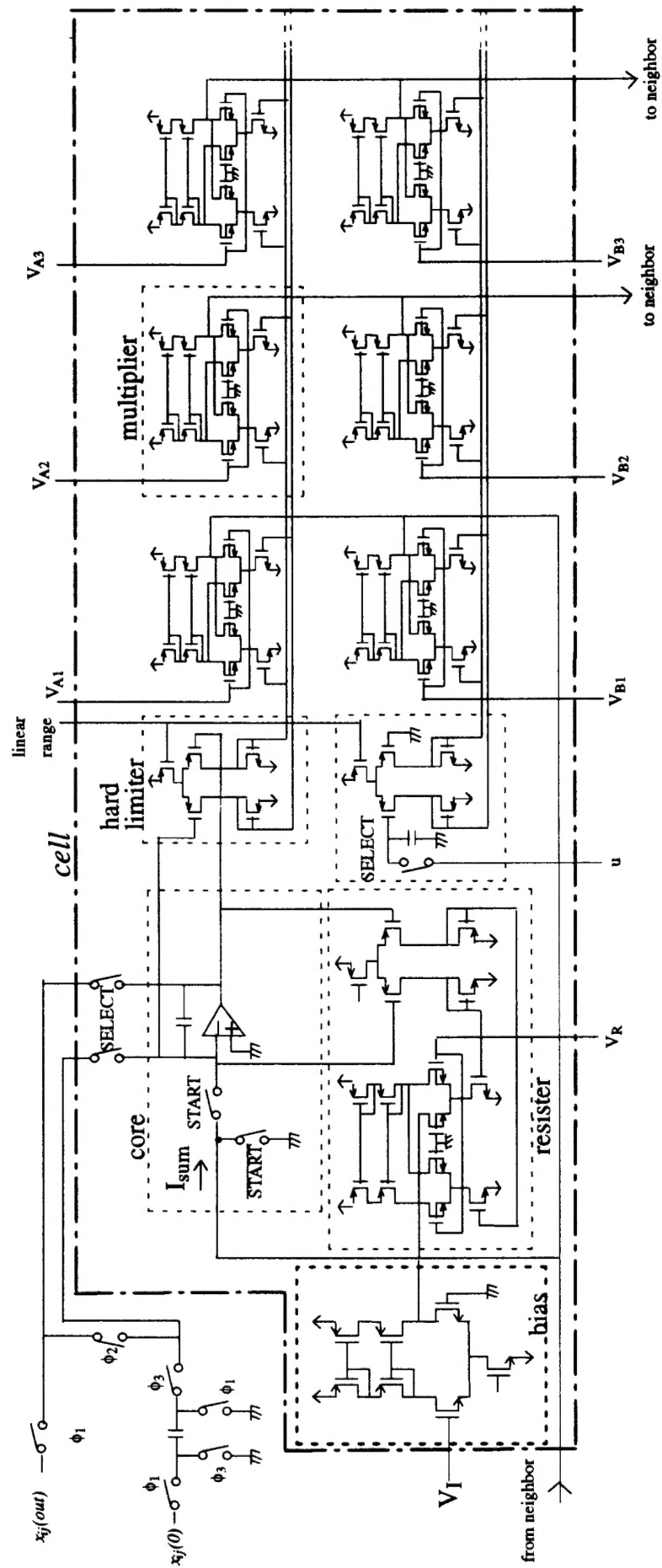
REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE
May 17, 1995 | 3. REPORT TYPE AND DATES COVERED
Technical Report 2/10/95 - 5/01/95 |
|---|---|---|

| 4. TITLE AND SUBTITLE
CNN VLSI Implementation | 5. FUNDING NUMBERS
G
N99914-94-1-0516 |
|---|---|

6. AUTHOR(S)

Jose Pineda de Gyvez

| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Texas Engineering Experiment Station
Texas A&M University | 8. PERFORMING ORGANIZATION REPORT NUMBER
93-549/#5 |
|---|---|

| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Code 214: JWK
Ballston Tower One
800 North Quincy Street
Arlington, Virginia 22217-5660	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

| 12a. DISTRIBUTION/AVAILABILITY STATEMENT
A: Approved for public release: distribution unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This report presents a CNN monolithic implementation suitable for Very Large Scale Integration. Novel circuit techniques have been investigated, designed and tested for various basic building blocks such as: *i)* a transconductance multiplier, *ii)* a linear resistor based on a multiplier circuit, *iii)* an OTA based integrator, *iv)* a hard limiter circuit and *v)* I/O circuitry based on Switched Capacitor Techniques. The report addresses both circuit design and layout plan issues. It highlights also the trade-offs among different design parameters including power supplies, dynamic range, area and tolerance. The CNN cell consumes approximately 2.5mW of power and occupies around $0.25mm^2$ of area. The IC has been sent for fabrication through MOSIS. Preliminary experimental results of some basic building blocks are encouraging.

| 14. SUBJECT TERMS
Cellular Neural Networks, VLSI | 15. NUMBER OF PAGES
19 |
|---|---|
| | 16. PRICE CODE |

| 17. SECURITY CLASSIFICATION OF REPORT
Unclassified | 18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified | 19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|